Infrared Physics and Technology 102 (2019) 103007

Contents lists available at ScienceDirect

Infrared Physics & Technology

journal homepage: www.elsevier.com/locate/infrared

Regular article

A novel FTIR analysis method for rapid high-confidence discrimination of esophageal cancer

James Ingham[a], Michael J. Pilling[b], David S. Martin[a], Caroline I. Smith[a], Barnaby G. Ellis[a], Conor A. Whitley[a], Michele R.F. Siggel-King[a], Paul Harrison[a], Timothy Craig[a], Andrea Varro[c], D. Mark Pritchard[c], Akos Varga[c], Peter Gardner[b], Peter Weightman[a], Steve Barrett[a,*]

[a] Department of Physics, University of Liverpool, Liverpool, United Kingdom
[b] Manchester Institute of Biotechnology, University of Manchester, Manchester, United Kingdom
[c] Department of Cellular and Molecular Physiology, Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom

ARTICLE INFO

Keywords:
Fourier transform infrared spectroscopy
Biomedical imaging
Biomarkers
Discrimination
Algorithm

ABSTRACT

It is demonstrated that a novel multivariate analysis technique can discriminate with accuracies in the range 81–97% between Fourier transform infrared (FTIR) images of esophageal cancer OE19 and OE21 cell lines, and between esophageal cancer associated myofibroblast (CAM) and adjacent tissue myofibroblast (ATM) cells. The latter cells are morphologically indistinguishable but are known to have functionally important differences in their capacity to stimulate cancer cell growth; this report provides the first accurate spectral discrimination between CAM and ATM cells taken from the same patient. Rapid and accurate discrimination between cell types was achieved, and key wavenumbers were identified which uniquely discriminate between all four cell types. This metrics-based analysis (MA) method is shown to be unique for distinguishing between cancer stromal cells from the same patient. The key wavenumbers differ significantly from those typically found to discriminate between various esophageal cell and tissue types. A comparison is made between the MA and the established Random Forest method, and the advantages of the MA are discussed. Crucially the findings suggest a novel method that allows cancer staging based discrimination of the stromal cell types that provide the niche for tumor development.

1. Introduction

Esophageal cancer is the sixth most common cause of cancer mortality [1–3] and is the cancer with the fastest rise in incidence in the western world. There are two main forms of esophageal cancer. One is squamous cell carcinoma, which is most common in Asia and is associated with smoking and poor diet. The other is adenocarcinoma, which is more common in the west and is associated with the gastro-esophageal reflux of acid and bile salts and the preneoplastic condition of Barrett's metaplasia of the esophagus [4,5]. Both cancers consist of malignant epithelial cells and stroma and the latter is important for facilitating cancer progression. One of the most important cell types in the stroma is a specialized fibroblast called the myofibroblast that produces growth factors and cytokines that promote cancer growth and metastasis [6,7]. The diagnosis of esophageal cancer follows the standard approach of examining images of excised tissue, obtained by endoscopy, after staining with Haematoxylin and Eosin (H&E). This highlights the nucleic acid and protein content of the specimen at blue and red visible wavelengths respectively. Typically, the interobserver discordance for the diagnosis of the low-grade dysplasia, which is characteristic of the earliest preneoplastic stage of disease is greater than 50% [8]. Although this discordance is reduced to ~15% for the diagnosis of the more serious condition of high-grade dysplasia, there is a need to improve the accuracy of diagnosis since false positives can give rise to unnecessary procedures and false negatives can be fatal [8–13]. As with all cancers, early detection is critical for the best patient outcome and there is a need for cheaper, more accurate and ideally automated methods for cancer diagnosis and for identification of those patients with Barrett's esophagus at most risk of progressing to dysplasia and cancer.

It has long been recognized that expanding the wavelength range of images of tissue will convey more information and there has been considerable progress in the application of infrared (IR) techniques to the examination of tissue in order to exploit the association of particular IR wavelengths with specific chemical moieties. Fourier transform infrared (FTIR) spectroscopy is one of the most successful techniques

* Corresponding author.
E-mail address: S.D.Barrett@liverpool.ac.uk (S. Barrett).

https://doi.org/10.1016/j.infrared.2019.103007
Received 2 May 2019; Received in revised form 5 August 2019; Accepted 5 August 2019
Available online 06 August 2019

applied to studies of cancer and has shown considerable promise for development into a diagnostic tool [14–19]. In particular there have been a number of previous applications of FTIR spectroscopy to the study of normal and cancer associated esophageal tissues [8–13]. Wang et al. [8] applied a partial least-squares fitting procedure to determine the principal components of the FTIR spectra of squamous, Barrett's non-dysplasia, Barrett's dysplasia and gastric tissue and Maziak et al. [9] gave a direct comparison of the FTIR spectra of normal and cancerous tissue. Quaroni and Casson [10] combined confocal FTIR microscopy and an analysis of second derivative FTIR spectra to distinguish normal and Barrett's esophageal tissue from adenocarcinoma. Amrania et al. [11] have developed 'Digistain', an instrument for use in histopathology that simplifies the analysis of IR spectra by comparing the intensity of two spectral features. Recently Old et al. [12,13] have developed an automated analysis technique for rapid IR mapping that identifies Barrett's dysplasia or adenocarcinoma with high sensitivity and specificity. The conclusions of this previous work are discussed in detail later.

Imaging FTIR typically yields information at each pixel in a two-dimensional image at ~1000 wavelengths, with each spectrum containing information on the many excitation modes of the large number of different molecular species contained in the specimen. Most reported work has analyzed these large data sets using techniques such as principal component analysis and the identification of 'fingerprints' for characterizing specimens, rather than at the level of detailed assignments of individual vibrational modes that is possible with simpler molecular systems.

In this paper the results of applying a novel multivariate analysis technique to FTIR spectra are described for two esophageal cancer cell lines, OE19 and OE21, and two esophageal myofibroblast cell lines derived from the stroma of an esophageal adenocarcinoma patient. OE19 was derived from an adenocarcinoma from the esophago-gastric junction and OE21 was derived from a squamous cell esophageal cancer. Both were purchased from HPA Culture Collections [20] and maintained as described previously [6,7]. The two myofibroblast cell lines were cancer associated myofibroblasts (CAM) and adjacent tissue myofibroblasts (ATM) obtained from the same patient and previously characterized [6,7].

It is now well recognized that tumor formation requires not just the acquisition of DNA mutations by cancer cells but also an appropriate cellular microenvironment (the cancer cell niche) that facilitates tumor growth and metastasis. Different stromal cell types are implicated in niche formation including inflammatory and immune cells, microvascular cells and cells of fibroblastic lineages. Myofibroblasts are an important sub-set of fibroblasts; CAMs are morphologically similar to ATMs that have been obtained from normal tissue adjacent to the cancer, but they differ markedly in their biology and in particular are strong stimulants of aggressive behaviors by cancer cells [6,21]. Transcriptomic, proteomic and miRNA profiling studies have all provided a basis for understanding the functional differences between CAMs and ATMs [6,22,23]. However, there remains a pressing need for methods that allow the rapid and precise identification of these cell types, not least because this would facilitate the identification of the cellular microenvironments in which tumor formation occurs.

The analysis technique described in this paper is able to discriminate between all four cell types with high accuracy and speed. This is particularly important for CAM and ATM cells in view of the much stronger capacity of the former in stimulating cancer cell growth and invasion [21–24]. The data therefore support the feasibility of new staging methods for early tumor development based on identifying the presence of those myofibroblasts (CAMs) most likely to facilitate cancer cell growth. Since early diagnosis improves patient outcomes this approach should bring clear benefits.

2. Materials and methods

Experiments were conducted on two esophageal cancer cell lines (OE19 and OE21) and two esophageal myofibroblast cells lines denoted CAM (cancer associated) and ATM (adjacent tissue associated). The CAM and ATM cells were obtained from the same patient undergoing surgery for esophageal adenocarcinoma [6,7]. This work was approved by the Ethics Committee of the University of Szeged, Hungary. Primary myofibroblast cultures were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 1% penicillin–streptomycin, 1% antibiotic–antimycotic and 1% non-essential amino acid solution as described previously [25] The OE19 and OE21 human Caucasian esophageal cells were obtained from HPA Culture Collections (Sigma, Dorset, UK) [20].

OE19 and OE21 cells were cultured at 37°C in a 5% CO_2 atmosphere in Roswell Park Memorial Institute (RPMI 1640) growth media (Sigma) supplemented with 2 mM glutamine (Sigma), 10% v/v fetal bovine serum (FBS) (Invitrogen, Paisley, UK) and 1% v/v penicillin/streptomycin (Sigma) until they reached 70–80% confluence. The culture medium was replenished at two-day intervals. The myofibroblast cells were cultured at 37°C in a 5% CO_2 atmosphere in Dulbecco's modified Eagle medium with L-glutamine containing 10% v/v FBS, 1% v/v modified Eagle medium nonessential amino acid solution, 1% v/v penicillin/streptomycin, and 2% antibiotic–antimycotic. Medium was replaced routinely every 48–60 h and cells were passaged at confluence, up to 12 times. CaF_2 discs (20 mm diameter × 2 mm thick, Crystran Ltd, Poole, UK) were sterilized using ethanol and rinsed with ultra-pure water and left to air-dry overnight. The discs were irradiated with UV for 30 min to ensure sterility. The sterile discs were then placed in each well of a tissue culture twelve-well plate (Corning, New York, USA). The cells (2×10^4 ml^{-1}) were seeded on each disc and incubated in a 5% CO_2 incubator at 37°C for two-days. After two-days the media was removed and the cells were fixed with a 4% v/v paraformaldehyde (PFA) (Sigma) solution and stored in 1x phosphate buffered saline (PBS) solution at 4 °C until required. Prior to imaging the CaF_2 slide containing the fixed cells was rinsed at least three times with Millipore ultra-pure water (18 MΩ cm). The rinsed slide was then removed from the well plate, the back surface wiped with ultra-pure water to ensure complete removal of any phosphate residue and then left to dry in the slide holder for a minimum of 90 min.

FTIR studies of the cell lines were carried out at room temperature in transmission mode with a Varian Cary 670-FTIR spectrometer in conjunction with a Varian Cary 620-FTIR imaging microscope produced by Varian (now Agilent Technologies, Santa Clara CA, USA) with a 128 × 128 pixel mercury-cadmium-telluride (MCT) focal plane array with a pixel size of 5.5 μm. The spectra were corrected for atmospheric and substrate absorption and the efficiencies of individual pixels in the array. FTIR images were acquired with a spectral range from 990 cm^{-1} to 3800 cm^{-1} with a resolution of 2 cm^{-1}, co-adding 256 scans. Infrared spectra were initially pre-processed using a principal component analysis based noise reduction algorithm. Substantial improvements in signal-to-noise were observed by retaining 10 principal components without the loss of biologically significant information. Spectra were then quality checked to remove those not attributable to the cell (including blank regions of the sample) or to a high degree of scattering. The quality check utilized a threshold based on the height of the Amide I band with spectra having absorbance between 0.03 and 1.00 being retained. Finally, infrared spectra were corrected for resonant Mie scattering with the RMieS-ESMC algorithm using 80 iterations and a matrigel reference spectrum [26–29].

3. Results

3.1. Data analysis method

An FTIR data cube was acquired for each cell type and was corrected

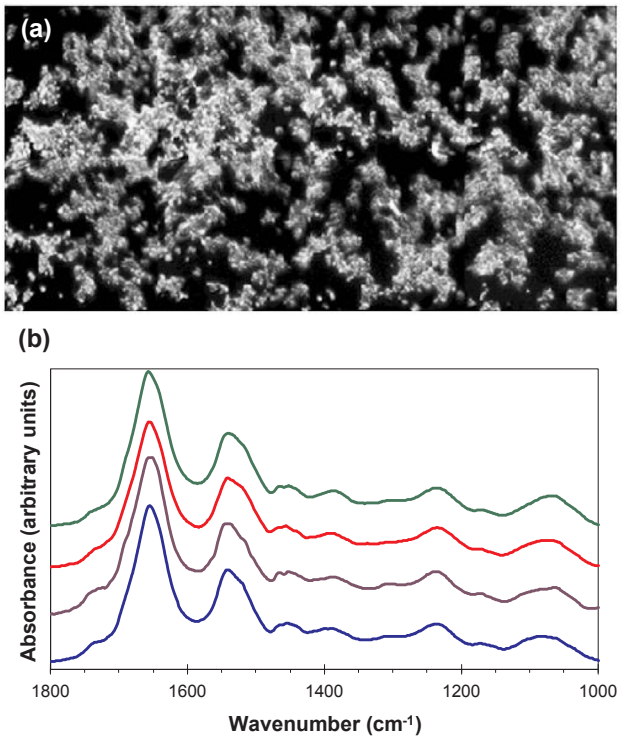

Fig. 1. (a) FTIR image of OE19 cells (~5000) integrated over all wavelengths, (b) the average FTIR spectra over all pixels for OE19 (green line), OE21 (red line), CAM (purple line) and ATM (blue line). The image is 2.8 mm × 1.4 mm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for Mie scattering effects [26]. Each FTIR data cube comprises a set of images of i × j pixels, where typically i × j ~10^5 and on average ~50% of pixels pass the quality check and Mie scattering correction. The third dimension of the data cube is the FTIR spectra of ~1400 data points covering the range of wavenumbers ν = 990 cm^{-1} to 3800 cm^{-1} in 2 cm^{-1} steps. The FTIR image obtained from the OE19 sample is shown in Fig. 1(a). The FTIR spectra characterizing each cell type [Fig. 1(b)] over the "fingerprint region" of 1000 cm^{-1} to 1800 cm^{-1}, were generated from averaging the spectra obtained from each pixel in the corresponding FTIR image of that cell type. This average does not include pixels from blank areas of the image. There are problems in deducing information from a direct comparison of these average profiles. Firstly, due to variations in the total intensity of the spectra obtained from each specimen, it is necessary to normalize each profile to the same area under the curve. Since the effect of the normalization on the spectral profile depends on the wavelength range used this can hide or exacerbate differences between the profiles of different specimens. Secondly, the standard deviation of the absorbance of all pixels at a given wavenumber is significant and shows significant overlap between cell types (see Supplementary Fig. 1). Consequently a more sophisticated analysis is required to reveal the differences between the spectral profiles of the different cell lines. There is considerable interest in the application of machine learning algorithms and multivariate analysis techniques to such problems and there are several recent reviews of the application of such techniques to FTIR spectra [30,31]. In this work a novel multivariate analysis method hereafter referred to as Metrics Analysis (MA) is described. The metrics were chosen to be the ratios of the absorbance for a given pair of wavenumbers. One advantage of this approach is that the results are independent of absolute absorbance and thus insensitive to factors such as sample thickness or normalization of the spectra. Importantly, this MA method treats all the data equally and does not attribute any biological significance to any particular wavenumber, in contrast to other work such as Fernandez et al. [32,33] in which discrimination of prostate tissues used metrics that were defined to have a significance related to tissue biochemistry. By examining ratios at wavenumbers over the whole range of 1000 cm^{-1} to 1800 cm^{-1}, the MA demonstrates the existence of biomarkers at wavenumbers that have not been identified in previous studies using other analysis techniques.

The MA method can be divided into three main parts: Stage 1: Training, Stage 2: Testing, and Stage 3: Analysis. For the results reported here, training was completed using 75% of the number of spectra in the data set, which were chosen at random, and testing was undertaken on the remaining 25%. Stage 1 parameterizes each cell type via the calculation of the absorbance ratio at two wavenumbers - the metric. This was done for all wavenumber combinations at a chosen step size over the range 1000 cm^{-1} to 1800 cm^{-1}. The step size was 6 cm^{-1}, as anything smaller has been shown [34] to be unnecessary. As a consequence there are a total of ~18000 metrics. In Stage 2 a score was then associated with each metric to quantify how well the metric was able to discriminate between cell types. For each cell type, scores were calculated by making distribution histograms for the metrics (one for the cell type and one for each of the other cell types in the analysis) where a high score is obtained for distributions that are distinct and hence have relatively little overlap. The score is defined by

$$\text{score} = \text{success rate} \times (1-\text{mislabeling rate})^2$$

where the success rate (often referred to as the sensitivity) is the rate at which the cell type is labeled correctly and the mislabeling rate (often referred to as the false positive rate) is the rate at which other cell types are labeled incorrectly as this cell type. Given that for the 25% of spectra used in this testing phase, the cell type is known, a success rate can be calculated and the probabilities of identifying the other cell types are used to determine the mislabeling rate. The scores for each metric are used to rank the ability of that metric to distinguish a given cell type. Stage 3 determines the number of metrics that are needed by a voting system to give the best overall success rate for cell type discrimination. The overall success rate is plotted as a function of the number of metrics used which indicates the optimal number of metrics required to achieve the best discrimination.

3.2. Discrimination between cell types

The wavenumbers that the MA method finds to be most important for discrimination can be visualized in a plot of the metric scores against ν_1 and ν_2, hereafter referred to as a Butterfly Plot. Two such plots, for CAM and ATM, are shown in Fig. 2.

All possible metrics are shown in these plots. The color-bar scale ranges from the least important (blue) to most important (red) metrics for discrimination. For the CAM and ATM samples, very different behavior is seen in the Butterfly plots, which highlights the clear discrimination achieved between these two cell types. This is a significant result since histopathologists find it difficult to distinguish between these cell types using the current standard method of optical microscopy on H&E stained samples [22]. For CAM, high scoring metrics are those that contain at least one high wavenumber around 1750 cm^{-1} (the red regions in Fig. 2(a)). The opposite situation is found for ATM, where high scoring metrics are often associated with at least one low wavenumber around 1150 cm^{-1} (the red regions in Fig. 2(b)).

While the scores for all the possible metrics (at the chosen step size) are evaluated and shown in Fig. 2, further insight can be obtained by limiting the results to a visualization of the best (highest-scoring) 100 metrics, hereafter referred to as Manhattan Plots. The plots for CAM and ATM are shown in Fig. 3, where the highest-ranked metrics for each cell type are shown plotted for ν_1 (red) and ν_2 (blue). These plots illustrate the combinations of wavenumbers that are used as a function of an increasing number of metrics from 1 to 100. It is clear that there are significant differences in the wavenumbers used for discrimination between these two cell types.

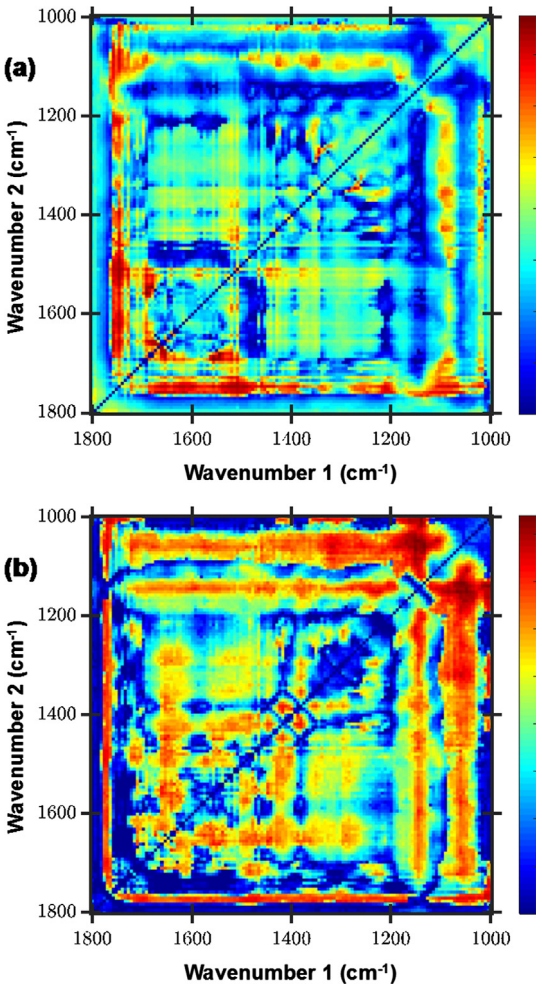

Fig. 2. Butterfly plots showing metric scores against wavenumbers ν_1 and ν_2 for (a) CAM and (b) ATM cell lines. Red (blue) indicates a relatively high (low) score. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In addition to visualizing the metric scores by Butterfly and Manhattan Plots, the success rate can be presented in a plot (Fig. 4) that shows how, for each cell type, the success rate varies with increasing number of metrics used in the analysis. In general, the success rate will eventually diminish due to poor metrics being added that compromise the success rate. Different variation is seen for the different cell types. For example, the success rate for ATM increases with the number of metrics used up to 24 metrics and subsequently decreases. In contrast, the success rate for OE19 is high for a low number of metrics and decreases as more metrics are used. For each cell type, the optimum number of metrics required for discrimination is given by the position of the maximum success rate.

As the data were sampled from a single image for each cell line, there was concern over whether spectra from adjacent pixels, which may be correlated due to the finite spatial resolution of the imaging system, could potentially bias the analysis and hence result in unrealistically high scores. To check this, the spatially ordered spectra were split into training and testing sets in such a way that the vast majority of the training spectra were not adjacent to the testing spectra. This analysis returned results that were indistinguishable from the original sets, demonstrating that any such pixel correlations do not contribute any significant bias to the results.

To aid the interpretation of the wavenumbers that are found to be important in this analysis, the wavenumbers in the top five metrics were examined for each cell type. Five metrics were chosen to give an

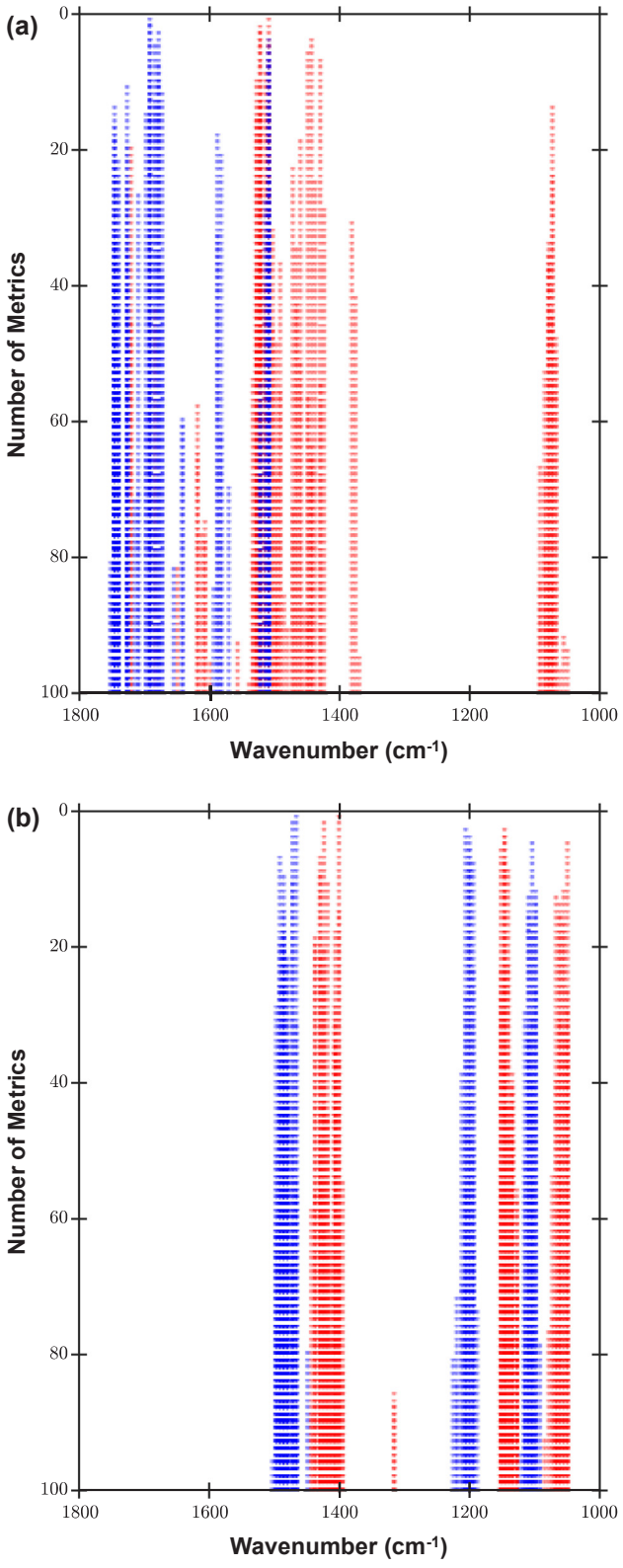

Fig. 3. Manhattan plots showing the combinations of wavenumbers, ν_1 (red) and ν_2 (blue), that are used as a function of an increasing number of metrics from 1 to 100 for (a) CAM and (a) ATM cell lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

apposite number of wavenumbers to allow meaningful comparisons between values for different cell types. These wavenumbers are shown in Fig. 5 and summarized in Table 1, and will be discussed further in the

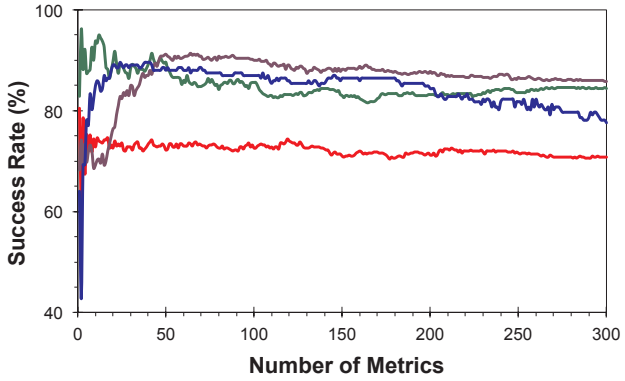

Fig. 4. Success Rate plot for each cell type, the optimum number of metrics required for discrimination of OE19 (green line), OE21 (red line), CAM (purple line) and ATM (blue line) are given by the position of the maximum success rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Discussion section.

3.3. Comparison of metrics analysis with random forest

In order to compare the MA method with existing classification methods we chose a quantitative comparison with the well-established random forest (RF) method. This is the most appropriate comparison as RF encapsulates both feature extraction and classification, and is commonly used for FTIR data analysis in the biomedical field. The same data sets were analyzed using both techniques for the four cell lines. The RF method used was a standard RF classification algorithm [35] available from https://github.com/tingliu/randomforest-matlab that was used to construct a classifier to discriminate between the different samples. Table 2 compares the MA and RF analysis results for the cell lines. The key wavenumbers found to be necessary for discrimination in both techniques showed some similarities. Little improvement in accuracy was seen when running the RF analysis for greater than ~30 s or by increasing the number of trees from 10 to 500. In general the MA method achieves greater accuracy in discrimination (particularly for ATM) in a shorter time (Table 2) than RF. For example, the MA of OE21 achieves a success rate of 79% within one minute whereas RF is limited to ~50%. It appears that RF is unable to distinguish ATM, with success rates no higher than would be expected from random chance (25%) when choosing one cell type from four possible types. These low success rates for the RF method are a consequence of the size of the data sets (the number of spectra) associated with each of the cell lines. The MA method gives high success rates regardless of whether the data sets are balanced and of comparable sizes, whereas the RF method is sensitive to this balance and gives poor success rates unless the data sets are rebalanced or the input data are reweighted.

4. Discussion

There have been significant advances in the application of FTIR to the study of normal and cancerous esophageal tissues [8–13]. Maziak et al. [9] compared FTIR profiles of normal and cancerous tissue and revealed prominent absorption changes at certain wavenumbers. In particular, changes at 964 cm^{-1} and 1237 cm^{-1} were assigned to increased nucleic acid content in malignant tissue, and changes in the bands at 1024 cm^{-1} and 1049 cm^{-1} indicated that glycogen was clearly present in healthy tissue but almost completely depleted in cancerous tissue [9]. Wang et al. [8] showed using a partial least-squares fitting procedure that the principal components of the FTIR spectra of squamous, Barrett's non-dysplasia, Barrett's dysplasia and gastric tissue in the range 950 cm^{-1} to 1800 cm^{-1} arose from

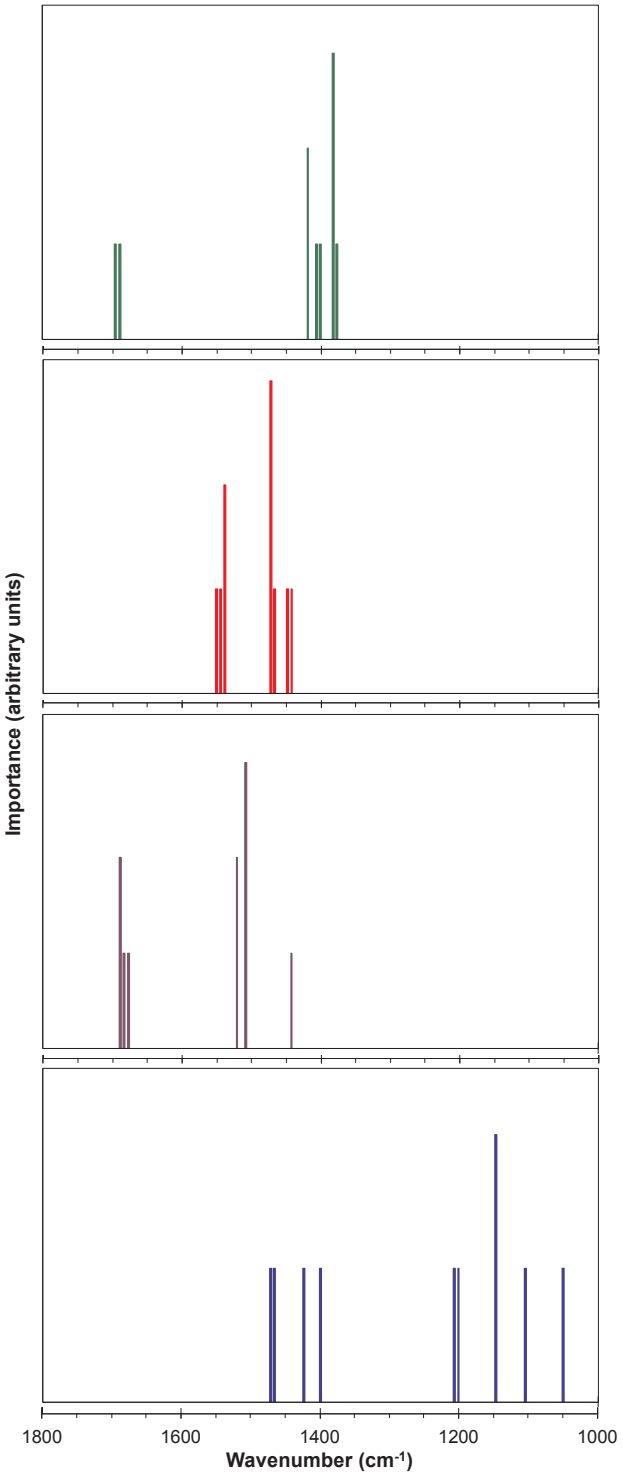

Fig. 5. Discrimination plot showing the histograms for the wavenumbers that are found to be important in discriminating between the four cell lines for the top five metrics: OE19 (green), OE21 (red), CAM (purple) and ATM (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variations in the concentration of DNA, protein, glycogen and glycoprotein. They established that dysplasia was characterized by an increase in glycoprotein and DNA. A subsequent imaging study by Quaroni and Casson [10] using a combination of confocal FTIR microscopy and a hierarchical cluster analysis of second derivative FTIR spectra was able to distinguish normal and Barrett's esophageal tissue

Table 1
Summary of Cell Line Metrics. Data on the optimum number of metrics, success rate at the optimum number of metrics and wavenumbers that discriminate between the four cell lines for the top five metrics.

Cell Type	Optimum Number of Metrics	Success Rate (%)	Wavenumbers for the top five metrics (cm^{-1})
OE19	2	97	1375, 1381, 1400, 1406, 1418, 1692, 1697
OE21	1	81	1443, 1449, 1466, 1472, 1539, 1545, 1551
CAM	64	92	1443, 1508, 1522, 1678, 1684, 1692
ATM	24	91	1049, 1103, 1146, 1200, 1206, 1400, 1424, 1466, 1472

Table 2
Comparison of MA and RF. Success rates (%) obtained by the metrics analysis (MA) and random forest (RF) approaches, for the cell lines.

	Random Forest		Metrics Analysis	
Number of trees	10	500	N/A	N/A
Resolution (cm^{-1})	20	20	20	6
Time (s)	27	1278	12	87
OE19 (%)	94	96	85	97
OE21 (%)	51	54	79	81
CAM (%)	94	96	83	92
ATM (%)	18	10	79	90
Mean of the four cell types (%)	64	64	81	90

from adenocarcinoma. They confirmed Wang et al.'s [8] association of glycoprotein bands with Barrett's and located these at the edge of crypts. Recently Old et al. [13] have developed a rapid IR mapping automated analysis technique that identifies Barrett's dysplasia or adenocarcinoma with 95.6% sensitivity and 86.4% specificity. Their analysis of second derivative FTIR spectra confirmed that normal squamous tissue had a high glycogen content, Barrett's tissue a high glycoprotein content and Barrett's dysplasia and adenocarcinoma a high DNA content.

The first thing to note from the results of the MA is that the wavenumbers that are found to discriminate between the different cell types (Table 1) differ significantly from the wavenumbers that have previously been used to characterize esophageal tissue types. For example, none of the glycogen, glycoprotein or DNA wavenumbers identified by Wang et al. [8] and Quaroni and Casson [10] or any of the ten characteristic wavenumbers identified in Table 7 of Old et al. [13] appear in Table 1. Also, only four of the twenty characteristic wavenumbers identified as distinguishing normal tissue from adenocarcinoma by Maziak et al. [9] appear in Table 1. This does not mean that the wavenumbers identified in previous work [8–10,13] are not valid discriminants (indeed, they are found by the MA when more metrics are included) but that they are not as significant as those found from the top five metrics.

The four wavenumbers common to this work and Maziak et al. [9] provide discriminants, to an accuracy of $\pm$ 1 cm^{-1}, of the following cells from all other cells; ATM (1049 cm^{-1}), OE19 and ATM (1399 cm^{-1}), OE19 and ATM (1465 cm^{-1}) and OE21 (1545 cm^{-1}). These wavenumbers are attributed, respectively, by Maziak et al. [9] to glycogen, lipids, lipids and proteins. The meaning of the wavenumbers found to discriminate between cell types in the MA is subtle since they are derived from a blind pair wise comparison of all the wavenumbers in the FTIR spectra of all the cell types. Consequently the discriminating wavenumbers must be interpreted with care. What is clear is that when used in combination with other metrics they provide excellent discrimination between all the cell types (Fig. 5). An analysis at the level of five metrics reveals twenty-four discriminating wavenumbers and as described in detail above, only four of these wavenumbers have been used in previous work to characterize differences between esophageal tissue types. Five of these discriminating wavenumbers in Table 1 are common to more than one cell type. A wavenumber that is common to two cell types means that it discriminates between those cells and all the others. This means that it is a characteristic of a chemical moiety that is either present or absent in those cells in a concentration that is significantly different to its concentration in all other cells.

The finding from previous work [8–10,13] that malignancy is characterized by an increase in DNA and a large decrease in glycogen suggests that changes in the concentration of these molecules should provide important discriminants between the ATM cells, which can be taken to be representative of healthy tissue, and the CAM cells and two malignant cell lines. This draws attention to the region between 1000 cm^{-1} and 1200 cm^{-1} where there is significant overlap between strong contributions from both molecules [9,36] and Table 1 and Fig. 5 show a strong concentration of discriminating wavenumbers in this spectral region. Fig. 6 shows an overlay of the normalized spectral profiles of Fig. 1 for each cell type in this spectral region. As explained earlier such comparisons of spectra can be misleading due to the dependence of the profiles on the wavelength range over which the normalization is carried out. However by taking a third power derivative of the spectra obtained from normal and malignant tissue Maziak et al. [9] identified four key wavenumbers in this region, 1024 cm^{-1}, 1049 cm^{-1}, 1080 cm^{-1} and 1155 cm^{-1} which they attributed to glycogen, glycogen, nucleic acids and proteins, respectively. Only one of these wavenumbers, 1049 cm^{-1}, occurs in the list of discriminating wavenumbers of Table 1 and Fig. 5. A deeper analysis of the data at the optimum number of metrics, twenty-four, reveals a large increase in the number of discriminating wavenumbers in this range as shown in Fig. 6. None of these additional wavenumbers correspond to the wavenumbers identified by Maziak et al. [9]. It is possible that some of the discriminating wavenumbers shown in Fig. 6 arise from particular chemical or structural effects in the DNA of the OE19, OE21 and CAM cell lines which could not be identified from tables of wavenumbers known to arise from particular chemical moieties.

A comparison of the other wavenumbers that discriminate between the different cell types and with the signatures of known chemical moieties [37,38] provides other insights into differences in chemical structure of the cells and tissues. For example the OE19 and CAM cells,

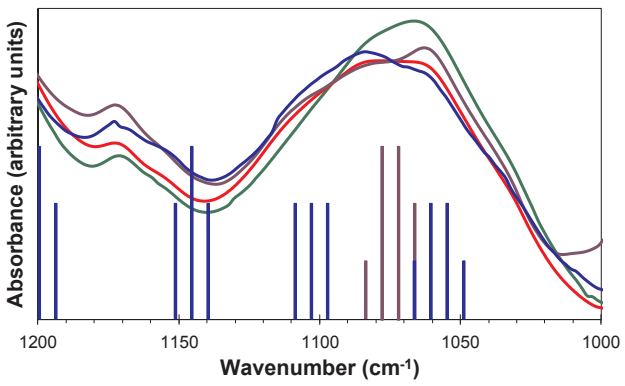

Fig. 6. Comparison of Spectral Profiles: the average spectra for the OE19 (green line), OE21 (red line), CAM (purple line) and ATM (blue line) cell lines in the region 1000–1200 cm^{-1} and histograms showing the wavenumbers that are found to be important in discriminating between the CAM (purple) and ATM (blue) cells for the optimum number of metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which are both derived from adenocarcinoma, share a discriminant at 1692 cm^{-1} associated with nucleic acids [37], which is absent from OE21 cells, which arise from squamous carcinoma. This wavenumber may be a moiety that is specific to adenocarcinoma. The OE21 and ATM cells share discriminating wavenumbers of 1466 cm^{-1} and 1472 cm^{-1}, which have been identified as characteristics of lipids [36–38].

It is particularly notable that the metrics approach provides excellent discrimination between cells derived from adenocarcinoma (OE19) and squamous cell carcinoma (OE21) and that ATM and CAM cells do not share a single one of the fifteen wavenumbers that discriminate between them and the other cell types. Clearly the identification of discriminating wavenumbers between the various cells types contain a wealth of information that is worthy of further study and may produce significant new insights into the chemical structure of esophageal and other cancers.

5. Conclusions

To summarize, we have demonstrated that a novel multivariate statistical analysis technique can discriminate with accuracies in the range 81% to 97% between FTIR images of OE19, OE21, CAM and ATM cell lines. This provides the first accurate spectral discrimination between CAM and ATM myofibroblast cells taken within 3 cm of tissue from the same patient. It should be stressed that these cell types are not readily distinguished by routine morphological approaches even though it is established that they have important biochemical differences that are relevant to the stimulation of cancer cell behavior [6]. The findings have potential clinical application in early diagnosis by identification of putative cancer cell microenvironments and by allowing the demarcation between tumor and adjacent tissue stroma without recourse to the analysis of biomarkers or extensive tissue processing. This is a significant result since histopathologists find it difficult to distinguish between these cell types using the current standard method of optical microscopy on H&E stained samples [22]. Moreover, the data indicate that it is now justified to conduct a much larger, appropriately powered, trial directed at the spectral discrimination of the important clinical groups, not least those Barrett's patients most at risk of progression including those with dysplastic lesions.

The MA method offers a new way of interpreting FTIR data. It has revealed wavenumbers which uniquely discriminate between all four cell types, many of which have not previously been identified with chemical moieties found in healthy tissue. The method discriminates between cells types with high accuracy and speed and has significant advantages over the RF approach. The method is expected to be widely applicable to other cell types and tissues.

Declaration of Competing Interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (UK EPSRC: Grant No. EP/K023349/1). JI, TC, BGE and CAW acknowledge support from EPSRC studentships. PG acknowledges the Williamson Trust for funds for the FTIR Imaging system.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.infrared.2019.103007.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA-Cancer J. Clin. 61 (2) (2011) 69–90.
[2] Y. Zhang, Epidemiology of esophageal cancer, World J. Gastroentero. 19 (34) (2013) 5598–5608.
[3] A. Pennathur, M.K. Gibson, B.A. Jobe, J.D. Luketich, Oesophageal carcinoma, Lancet 381 (9864) (2013) 400–412.
[4] B.J. Reid, X. Li, P.C. Galipeau, T.L. Vaughan, Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis, Nat. Rev. Cancer 10 (2) (2010) 87–101.
[5] T.K. Desai, K. Krishnan, N. Samala, J. Singh, J. Cluley, S. Perla, C.W. Howden, The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: a meta-analysis, Gut 61 (7) (2012) 970–976.
[6] J.D. Kumar, C. Holmberg, S. Kandola, I. Steele, P. Hegyi, L. Tiszlavicz, R. Jenkins, R.J. Beynon, D. Peeney, O.T. Giger, A. Alqahtani, T.C. Wang, T.T. Charvat, M. Penfold, G.J. Dockray, A. Varro, Increased expression of chemerin in squamous esophageal cancer myofibroblasts and role in recruitment of mesenchymal stromal cells, PloS ONE 9 (8) (2014) e104877.
[7] J.D. Kumar, S. Kandola, L. Tiszlavicz, Z. Reisz, G.J. Dockray, A. Varro, The role of chemerin and ChemR23 in stimulating the invasion of squamous oesophageal cancer cells, Brit. J. Cancer 114 (10) (2016) 1152–1159.
[8] T.D. Wang, G. Triadafilopoulos, J.M. Crawford, L.R. Dixon, T. Bhandari, P. Sahbaie, S. Friedland, R. Soetikno, C.H. Contag, Detection of endogenous biomolecules in Barrett's esophagus by Fourier transform infrared spectroscopy, Proc. Natl. Acad. Sci. USA 104 (40) (2007) 15864–15869.
[9] D.E. Maziak, M.T. Do, F.M. Shamji, S.R. Sundaresan, D.G. Perkins, P.T. Wong, Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: an exploratory study, Cancer Detect. Prev. 31 (3) (2007) 244–253.
[10] L. Quaroni, A.G. Casson, Characterization of Barrett esophagus and esophageal adenocarcinoma by Fourier-transform infrared microscopy, Analyst 134 (6) (2009) 1240–1246.
[11] H. Amrania, G. Antonacci, C.-H. Chan, L. Drummond, W.R. Otto, N.A. Wright, C. Phillips, Digistain: a digital staining instrument for histopathology, Opt Exp. 20 (7) (2012) 7290–7299.
[12] O. Old, G. Lloyd, M. Isabelle, L.M. Almond, C. Kendall, K. Baxter, N. Shepherd, A. Shore, N. Stone, H. Barr, Automated cytological detection of Barrett's neoplasia with infrared spectroscopy, J. Gastroenterol. 53 (2) (2018) 227–235.
[13] O.J. Old, G.R. Lloyd, J. Nallala, M. Isabelle, L.M. Almond, N.A. Shepherd, C.A. Kendall, A.C. Shore, H. Barr, N. Stone, Rapid infrared mapping for highly accurate automated histology in Barrett's oesophagus, Analyst 142 (8) (2017) 1227–1234.
[14] E. Gazi, J. Dwyer, N. Lockyer, P. Gardner, J.C. Vickerman, J. Miyan, C.A. Hart, M. Brown, J.H. Shanks, N. Clarke, Application of FTIR Microspectroscopy and ToF-SIMS Imaging in the Study of Prostate Cancer, Faraday Discussions 126 (2004) 41–59.
[15] K. Gajjar, J. Trevisan, G. Owens, P.J. Keating, N.J. Wood, H.F. Stringfellow, P.L. Martin-Hirsch, F.L. Martin, Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer, Analyst 138 (14) (2013) 3917–3926.
[16] B.R. Smith, K.M. Ashton, A. Brodbelt, T. Dawson, M.D. Jenkinson, N.T. Hunt, D.S. Palmer, M.J. Baker, Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology, Analyst 141 (12) (2016) 3668–3678.
[17] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst 142 (8) (2017) 1258–1268.
[18] S. Kumar, C. Desmedt, D. Larsimont, C. Sotiriou, E. Goormaghtigh, Change in the microenvironment of breast cancer studied by FTIR imaging, Analyst 138 (14) (2013) 4058–4065.
[19] P. Bassan, J. Mellor, J. Shapiro, K.J. Williams, M.P. Lisanti, P. Gardner, Transmission FT-IR chemical imaging on glass substrates: applications in infrared spectral histopathology, Anal. Chem. 86 (3) (2014) 1648–1653.
[20] J.C. Rockett, K. Larkin, S.J. Darnton, A.G. Morris, H.R. Matthews, Five newly established oesophageal carcinoma cell lines: phenotypic and immunological characterization, Brit. J. Cancer 75 (2) (1997) 258–263.
[21] O. De Wever, M. Van Bockstal, M. Mareel, A. Hendrix, M. Bracke, Carcinoma-associated fibroblasts provide operational flexibility in metastasis, Semin. Cancer Biol. 25 (2014) 33–46.
[22] C. Holmberg, M. Quante, I. Steele, J.D. Kumar, S. Balabanova, C. Duval, M. Czepan, Z. Rakonczay Jr, L. Tiszlavicz, I. Nemeth, G. Lazar, Z. Simonka, R. Jenkins, P. Hegyi, T.C. Wang, G.J. Dockray, A. Varro, Release of TGFbetaig-h3 by gastric myofibroblasts slows tumor growth and is decreased with cancer progression, Carcinogenesis 33 (8) (2012) 1553–1562.
[23] L. Wang, I. Steele, J.D. Kumar, R. Dimaline, P.V. Jithesh, L. Tiszlavicz, Z. Reisz, G.J. Dockray, A. Varro, Distinct miRNA profiles in normal and gastric cancer myofibroblasts and significance in Wnt signaling, Am. J. Physiol-Gastr L. 310 (9) (2016) G696–G704.
[24] L. Jiang, T.A. Gonda, M.V. Gamble, M. Salas, V. Seshan, S. Tu, W.S. Twaddell, P. Hegyi, G. Lazar, I. Steele, A. Varro, T.C. Wang, B. Tycko, Global hypomethylation of genomic DNA in cancer-associated myofibroblasts, Cancer Res. 68 (23) (2008) 9900–9908.
[25] H. Najgebauer, T. Liloglou, P.V. Jithesh, O.T. Giger, A. Varro, C.M. Sanderson, Integrated omics profiling reveals novel patterns of epigenetic programming in

cancer-associated myofibroblasts, Carcinogenesis 40 (4) (2019) 500–512.
[26] P. Bassan, A. Kohler, H. Martens, J. Lee, H.J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, P. Gardner, Correction of infrared spectra from highly scattering biological samples, Analyst 135 (2) (2010) 268–277.
[27] H. Martens, E. Stark, Extended multiplicative signal correction and spectral interference subtraction - new preprocessing methods for near- infrared spectroscopy, J. Pharmaceut. Biomed. 9 (8) (1991) 625–635.
[28] A. Kohler, J. Sulé-Suso, G.D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D.G. van Pittius, G. Parkes, H. Martens, Estimating and correcting Mie scattering in synchrotron based microscopic FTIR spectra by extended multiplicative signal correction (EMSC), Appl. Spectrosc. 62 (3) (2008) 259–266.
[29] P. Bassan, H.J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, Resonant Mie scattering in Infrared spectroscopy of biological materials – understanding the "dispersion artifact, Analyst 134 (8) (2009) 1586–1593.
[30] T.P. Wrobel, R. Bhargava, Infrared spectroscopic imaging advances as an analytical technology for biomedical sciences, Anal. Chem. 90 (3) (2018) 1444–1463.
[31] S. Berisha, M. Lotfollahi, J. Jahanipour, I. Gurcan, M. Walsh, R. Bhargava, H. Van Nguyen, D. Mayerich, Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks, Analyst 144 (5) (2019) 1642–1653.
[32] D.C. Fernandez, R. Bhargava, S.M. Hewitt, I.W. Levin, Infrared spectroscopic imaging for histopathologic recognition, Nature Biotechnol. 23 (4) (2005) 469–474.
[33] R. Bhargava, D.C. Fernandez, S.M. Hewitt, I.W. Levin, High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data, Biochimica Biophysica. 1758 (7) (2006) 830–845.
[34] R. Bhargava, Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology, Anal. Bioanal. Chem. 389 (4) (2007) 1155–1169.
[35] L. Breiman, Random Forests, Mach. Learn. 45 (1) (2001) 5–32.
[36] L. Chiriboga, P. Xie, H. Yee, V. Vigorita, D. Zarou, D. Zakim, M. Diem, Infrared spectroscopy of human tissue. I. differentiation and maturation of epithelial cells in the human cervix, Biospectroscopy 4 (1998) 47–53.
[37] A.C.S. Talari, M.A.G. Martinez, Z. Movasaghi, S. Rehman, I. ur Rehman, Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 52 (5) (2017) 456–506.
[38] Z. Movasaghi, S. Rehman, I. ur Rehman, Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 43 (2) (2008) 134–179.